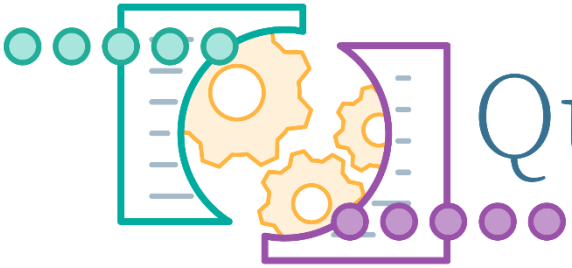


QuestionLink Engine

A methodological and technical Primer on the QuestionLink Engine

Ranjit K. Singh & Matthias Roth

Mannheim, Germany, 2022-01-24

Abstract

In this primer, we give an overview of the QuestionLink Engine; the R package with which we create the recoding documents for QuestionLink. We discuss the method we use to harmonize different survey instruments for the same construct: Observed Score Equating in a Random Groups design (OSE-RG). Since OSE-RG relies on samples of two instruments drawn randomly from the same population, we discuss how we can use data from surveys with probabilistic national samples to satisfy this random groups precondition. On the technological side, we explain how the QuestionLink Engine searches for such random groups instances, how it performs the possible equatings, how it aggregates them to the recoding tables that QuestionLink offers, and finally how it creates the interactive recoding documents. Lastly, this primer addresses some methodological and technological miscellanea, that shape the way the QuestionLink Engine works in its current form.

Table of contents

Introduction

Researchers increasingly often combine data from multiple surveys into one harmonized dataset for their research projects (e.g., Hussong et al., 2013; Schulz & Weiß, 2014; Slomczynski & Tomescu-Dubrow, 2018). Harmonization of survey data, however, comes with its own challenges. Frequently, social science surveys use single question measurement instruments to measure latent constructs, such as values, attitudes, or interest. Such single-question instruments often differ in question design characteristics such as the wording of their question text, of the response options, the number of response options or their layout (Menold & Bogner, 2016; Tourangeau et al., 2000). The variability of such single-question measurement instruments (in the following, just instruments) creates an obstacle to researchers wanting to compare or combine measurements of latent constructs across different surveys or even within the same survey if instruments were changed over time.

QuestionLink tackles this problem by creating recoding tables which translate the numerical values of an instrument B into the numerical form of a reference instrument A (Kolen & Brennan, 2014; Singh, 2020). Table 1 illustrates this for an example where we harmonize an instrument B with four ordinal response options (i.e., four scale points) towards the format of a reference instrument A with five scale points.

Table 1

A recoding table to transform instrument B into the format of instrument A

Possible responses in instrument B	Equivalents of instrument B responses in the format of instrument A	Possible responses in reference instrument A
1	1.63	1
2	2.89	2
3	3.92	3
4	4.89	4
—	—	5

If we recode responses to instrument B into their instrument A equivalents, we increase the comparability between measurements with instrument A and B. Specifically, we ensure that if we measured the same population, we would now get a similar response distribution regardless which instrument we used. This means we would get similar average response values, a similar standard deviation, skewness, and percentiles. Such a transformation allows us to compare results derived from both instruments and also to combine data measured with both instruments for joint analyses. Please note, that QuestionLink provides such tables for any combination of instruments and in both directions (A towards B and B towards A, for example).

The translation tables are provided in the form of recoding scripts which contain the recoding information for several major statistical software packages. The QuestionLink engine, meanwhile, is a package of R scripts which automates the process with which the recoding information is calculated and compiled in an easy-to-use format for end-users.

In this methodological and technical primer, we will give a quick introduction to the method at the heart of QuestionLink: Observed-Score Equating in a Random Groups Design (OSE-RG) using the equipercentile equating algorithm (Kolen & Brennan, 2014; Singh, 2020). We will then explain the logic and architecture of the QuestionLink engine, which automates large parts of the harmonization process. Lastly, we will discuss some technical details and decisions that the QuestionLink engine is currently based on.

Observed Score Equating in a Random Groups Design (OSE-RG)

The method at the heart of QuestionLink, OSE-RG, is part of a family of approaches from psychometry collectively called equating. Equating aims to make results of different measurement instruments for the same latent construct comparable (Kolen & Brennan, 2014; Price, 2017). The original context of equating is the harmonization of complex, multi-item tests in psychometric diagnostics (e.g., professional aptitude or mental ability tests). Many equating approaches thus make use of the multi-item structure of psychometric instruments.

OSE-RG, however, can be applied to single-question instruments as well. Consequently, it is most suitable to harmonizing survey data in the social sciences. Observed Score Equating (OSE) aims to increase comparability by ensuring that we get comparable response distributions for the same population regardless of the instrument used (Kolen & Brennan, 2014). This means that the arithmetic mean, the standard deviation, and higher distribution moments such as skewness are now more comparable than before. In practice, this means that respondents with a similar level of the measured construct (e.g., political interest) will on average be represented with the same numerical value after we have applied the recoding script.

The random groups (RG) design

How does OSE-RG achieve this? Here the random groups part of OSE-RG comes into play. The random groups design is comparable to a split-ballot experiment (i.e., a between-subject random experiment with two conditions): We apply the two instruments to random samples of the same population (Kolen & Brennan, 2014). The goal is to control true population differences in the construct distribution by setting the population equal and thus isolating differences in measurement. The two different instruments will result in two different response distributions. However, since both samples are randomly drawn from the same population, these response distribution differences represent differences in measurement. To give a concrete example: In two sufficiently large random samples of the same population, we would expect a very similar level of political interest. If the measured scores now differ between the two instruments in their respective samples, then these differences are most likely due to instrument differences. Consequently, we can increase comparability, if we now recode one instrument so that its response distribution shape matches that of the other instrument. This means, for example that an average person from our population is now represented with the same average value in the recoded data, regardless of the instrument used. And the same goes for above- and blow average persons as well.

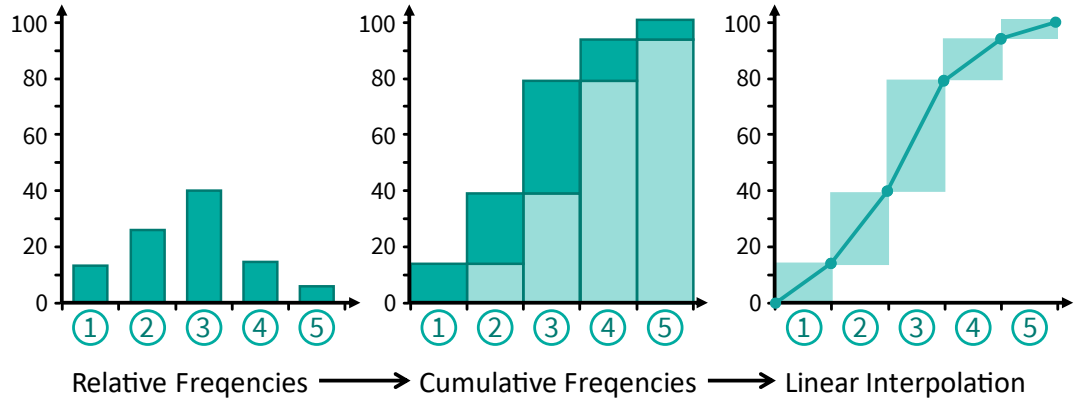
The equipercentile equating algorithm

The last issue is then how to align the response distribution shapes. There are different algorithms, but for QuestionLink we decided on the equipercentile equating algorithm. The advantage of this algorithm is that it can handle non-normal (e.g., very skewed) response distribution shapes with ease (Kolen & Brennan, 2014). Imagine two instruments for the same construct, A and B. Now imagine we have samples for both instruments randomly drawn from the same population. Equipercentile OSE-RG would then assign each response option in A and each response option in B a percentile rank. Response options would then be matched by their percentile rank. If 50% of respondents choose one response in A and another in B, we would match those. Consequently, after harmonization with OSE-RG, values of different instruments are comparable because they point to the same level of construct intensity as represented by the ordered position in a common population.

In practice, the equipercentile equating algorithm is not quite so simple, of course. The reason for this is that responses do not directly represent one specific percentile rank. Instead, they represent a range of percentiles in a population. If 14% choose the first response option in an instrument, then this response represents the 0^{th} to 14^{th} percentile rank for that population. Equipercentile equating solves this via linear interpolation (Kolen & Brennan, 2014). It assumes that all percentile ranks in such a range are equally likely. In visual terms, it draws a line from 0% to 14%; hence linear interpolation. This implies that we would assign the first response option in our example the middle of it ranges as its percentile (i.e., the 7^{th} percentile). Figure 1 below illustrates this process for a single-item instrument with five response options.

Figure 1

Linear interpolation of percentiles



Please note that this percentile interpolation process works in both directions. We can transform numerical response scores into percentiles and percentiles back into their corresponding response scores. In fact, we can transform arbitrary percentiles. If they do not perfectly match a specific response, we simply obtain a decimal number instead of an integer. A “1.5” means that the percentile is halfway between the first (1) and the second (2) response option.

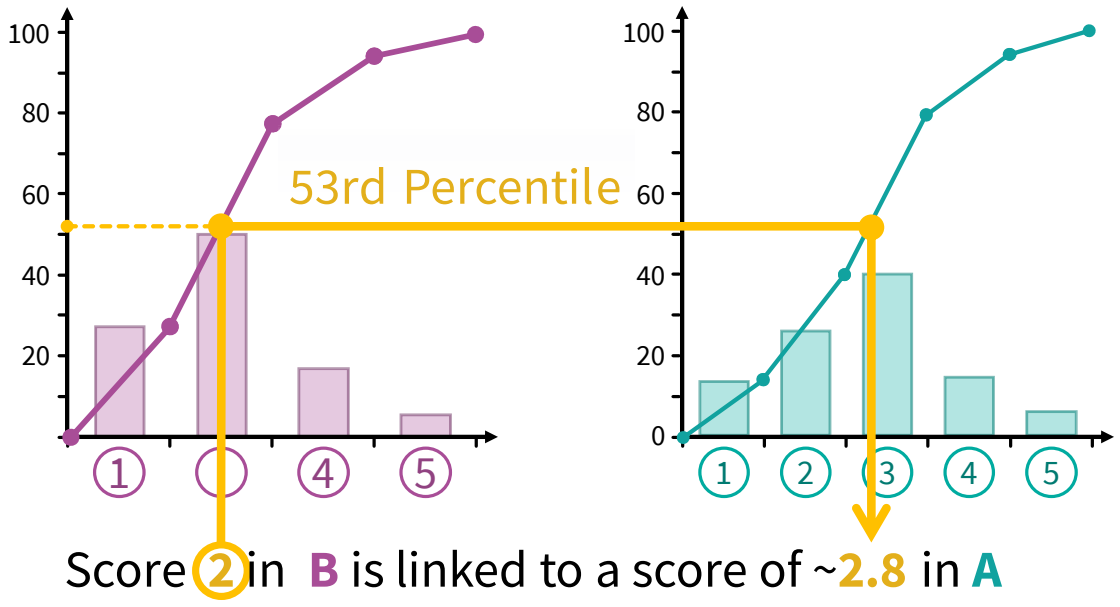
Thus, once we have interpolated the percentile ranks for both instruments, A and B, the rest of the process is simple. If we want to harmonize responses of instrument B towards a reference instrument A, then we would:

1. Transform the numerical responses to instrument B (i.e., the integer scores 1, 2, etc.) into their interpolated percentiles.
2. Then transform these percentiles into the respective, interpolated response scores in terms of the reference instrument A.

In figure 2 below, we see an example where instrument B with four response options is harmonized with the reference instrument A with five response options. Specifically, the example shows how first we transform a score of 2 in B (i.e., B's second response option) into its corresponding percentile in the random groups population (i.e., the 53rd percentile). Second, we then transfer this percentile back into a response score; this time one of instrument A. A percentile rank of 53 happens to correspond with a 2.8 in instrument A. Hence, we can now recode a score of 2 in instrument B into a harmonized score of 2.8 to match instrument A.

Figure 2

The equipercentile equating algorithm visualized



The end result is then a recoding (or correspondence) table in which every untransformed possible response in instrument B is matched with an OSE-RG harmonized equivalent value in terms of the reference instrument A. This recoding table can then be used in other instances where the same instruments were used as well.

Further reading:

For an easily accessible introduction into the underlying harmonization challenge and the benefits of OSE-RG I would recommend several of the illustrated posts in our GESIS Blog series on ex-post harmonization:

- *Ceci n'est pas une pipe: Disentangling measurement and reality in ex-post harmonization*
  An introduction into the basic problem of comparing different measurement instruments.
- *(Not) by any stretch of the imagination: A cautionary tale about linear stretching*
  A post explain why linear stretching, i.e., just matching minimum and maximum response options and stretching everything in between, is not sufficient.
- *The new normal: Linear equating of different instruments*
  A primer on OSE-RG using the easier to understand linear equating algorithm.
- *Cats are liquids: Equipercentile equating of different instruments*
  And finally, a step-by-step explanation of the equipercentile equating algorithm.

For a formal introduction to equating we recommend the very instructional book by Kolen and Brennan (2014).

- Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Automating the random groups design in QuestionLink

As we have seen, OSE-RG becomes possible if we have suitable data in a random groups design, that is random samples for both instruments from the same population. Obtaining suitable random groups data can be a challenge. QuestionLink thus overcomes this challenge by integrating data from several large, long-running surveys which all randomly sample the adult German population. To obtain suitable random groups data for OSE-RG, we thus have to find instances where two measurement instruments were applied to the adult German population in the same year (Singh, 2020). In such an instance we have all we need to harmonize the two instruments via OSE-RG.

However, is not quite that simple in practice for two reasons: Firstly, QuestionLink aims to harmonize all instruments of the integrated surveys with each other. Hence, we have to perform an increasing number of harmonizations, because the number of possible instrument pairs rises supralinearly. It is in essence the handshake problem, meaning for n instruments we get $\frac{(n-1)\cdot n}{2}$ instrument pairs. For five instruments we need to harmonize ten pairs, for ten instruments, we need to harmonize 45 pairs, and for fifteen instruments we need to harmonize 105 pairs. That alone calls for automation.

Secondly, we may not have instances with the same year for all instrument pairs. Applying OSE-RG when the samples for both instruments stem from the same country but from different years might lead to bias because the true population distribution might have changed over time. For example, Germans might have become more interested in politics after a decade.

QuestionLink thus uses three different approaches to linking instruments:

1. **Direct Links:**
   In a direct link we have data for both instruments from the same year (in the adult German population).
2. **Time-relaxed Links:**
   If the difference between the samples for both instruments is less than a predefined number of years, a time-relaxed link is formed. Currently, QuestionLink allows a maximum time-relaxation of one year (e.g., performing OSE-RG across data from 2000 and 2001).
3. **Relay Links:**
   If the instruments were used in years further than time-relaxation apart, then OSE-RG is applied via a relay instrument. OSE-RG, like all forms of equating, can be chained, meaning we can harmonize instrument A into instrument B via a relay instrument C. Relay links thus look for direct or time-relaxed links between instrument A and a relay and then such links from the relay towards the target instrument B. This might mean harmonizing A to a relay in 1988 and then harmonizing the relay further towards B in 2002.

Both time-relaxed and relay links have drawbacks compared to direct links. Time-relaxation might incur some bias if the true construct distribution changed drastically in Germany within a year. Relay linking meanwhile incurs random sampling errors and the resulting standard error of equating twice, once for each equating. However, both drawbacks are mitigated in QuestionLink, because the engine uses all available links simultaneously. The final recoding table is the median of all those links. Random fluctuations and temporal shifts tend to cancel out across so many links. For the seven measures for political interest, for example, the QuestionLink engine identified and processed more than 30.000 links which were then condensed down to 42 recoding tables.

QuestionLink Architecture

The QuestionLink engine is a collection of R functions which automates the process of generating recoding tables and scripts for a specific construct measured with different instruments in the survey programs include in QuestionLink. It furthermore creates easy-to-use documents for end-users which contain the recoding information as scripts for several major statistical software packages. However, QuestionLink requires some manual work to prepare response data and instrument information for processing.

The QuestionLink engine thus accepts some specifically formatted inputs and generates the interactive recoding documents which we then make available to users. Specifically, QuestionLink generates one recoding document for each defined instrument, because each document recodes towards that instrument's numerical format.

Required Input

The QuestionLink engine requires two inputs. The first input is a large, structured dataset which contains all responses to the instruments for a construct across all surveys. This dataset enriches each response with information on the instrument used, the survey it was collected in, the year it was collected in, sample weighting information, and some additional information listed later

under “Technical and Methodological specifics”. The second input is a documentation table which contains the question wording, the response option wording, and their respective English translations for each instrument. This table ensures that the final output has consistent documentation.

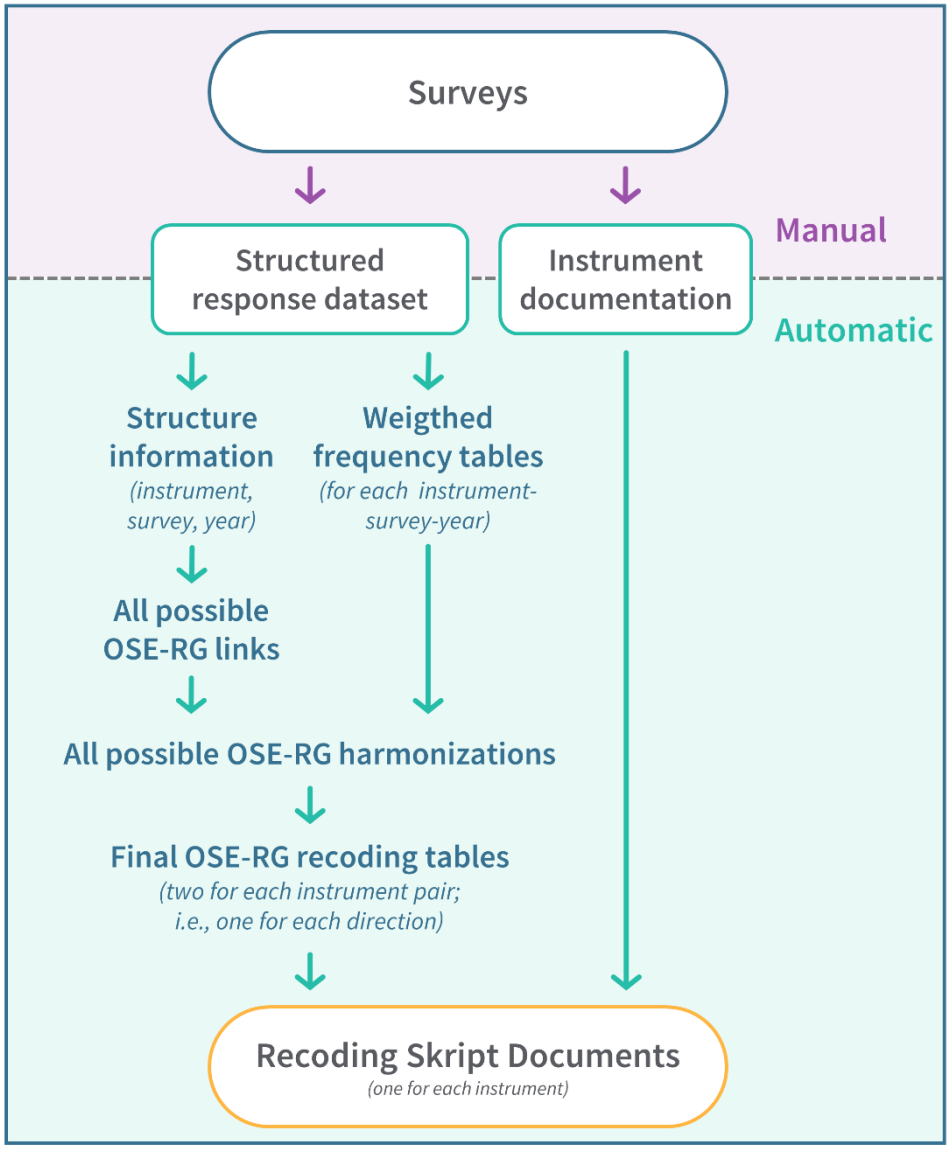
In short, the structured response dataset provides both the structure information (which instrument, which survey, what year) as well as the response information necessary for performing OSE-RG harmonization. The documentation table, meanwhile, is only referred to while generating the final recoding documents to ensure consistency. Based on the structured response data set QuestionLink fully automates the remaining equating process.

Automated equating procedure

The automated equating proceedure takes the manually prepared inputs and returns the recoding script documents intended for the user.

Figure 3

An overview of the QuestionLink process



The QuestionLink engine starts out by extracting structure information from the **structured response dataset** in which all responses across all surveys, instruments, and years are stored. The resulting **structure information** table details which instrument was used in which surveys and in which years. This information is crucial, because we want to find possible links for OSE-RG. That means instances where instrument pairs were used in the same year (direct link), in adjacent years (time-relaxation link), or instances where we can link two instruments via a third relay instrument temporally overlapping them both (relay link). Thus the QuestionLink engine derives **all possible OSE-RG links** for all intstrument pairs from the structure information. These links are in essence treated as a long to-do list of OSE-RG harmonizations to be performed. However, to apply OSE-RG, we require frequency tables for each instrument in each survey in each year. Thus, the engine also creates **weighted frequency tables** for every instrument–survey–year combination. The tables are weighted in the sense that the sourcer survey's design weights are applied when creating the frequency tables.

Now everything is in place to actually calculate all possible recoding tables. The QuestionLink engine takes every possible link between every instrument pair and performs an OSE-RG harmonization. To this end, it looks up the link, then fetches the relevant frequency tables, and finally feeds those into an equipercentile equating function. In case of a direct link, this might mean fetching the frequency table for instrument A in a specific year in a specific survey and then fetching the frequency table for instrument B in the same year in a specific survey. Then both frequency tables are processed using the equipercentile equating algorithm. Please note that the result is not one but two recoding tables. One to harmonize instrument A towards B, and one to harmonize instrument B to A. A time-relaxed link works the same way, except that the years for A and B are adjacent instead of identical. Lastly, relay links are more complex, because four frequency tables have to be fetched instead of two. One for instrument A, one for the relay in a year which overlaps with instrument A's years, another one for the relay instrument which overlaps with instrument B's years, and finally one for instrument B.

After this process, we have **all possible OSE-RG harmonizations** in the form of a datastructure containing all possible recoding tables between all instruments. Depending on the number of instruments and the years in which they were applied, the number of links and thus the number of recoding tables usually ranges in the tens of thousands (10^4). In principle, however, each separate harmonization of two instruments should result in a very similar recoding table. Still, harmonizations of the same two instruments but via different links (meaning different years, different surveys, or different link types) fluctuate due to different error sources. There is some degree of random error, for example because equating is based on random samples instead of population statistics. Also, time-relaxed links might be biased through changes in the true construct distribution between adjacent years. This is the reason why QuestionLink performs so many different OSE-RG harmonizations and then aggregates them to mitigate such errors. Specifically, QuestionLink derives the **final recoding tables** which users get to see as follows. For the case of harmonizing instrument B towards the format of instrument A, QuestionLink loops through every possible response in instrument B. For each possible response it fetches every OSE-RG harmonized equivalent in terms of the format of instrument A. Then the median equivalent is chosen as the final equivalent. The result is a recoding table where every possible response to

instrument B is transformed with the median equivalent in instrument A across all possible OSE-RG harmonizations.

This means that at this point in the process, we now have a new datastructure, which contains $n \cdot (n-1)$ recoding tables where n is the number of instruments for a construct that was entered into QuestionLink: Two for each possible instrument pair. All that remains now is to create the **recoding script documents** that users can download. For ease of use, we have decided to provide one recoding script document per instrument. The idea is that users can select any of the covered instruments as a reference instrument, meaning that all other instruments are harmonized towards that reference instrument's format. The QuestionLink engine creates these documents automatically via R Markdown. It can handle an arbitrary number of instruments and extends the document accordingly with modular components. QuestionLink loops through the relevant instrument pairs and their recoding tables and generates recoding script snippets for R, STATA, and SPSS automatically. It also gives information on the OSE-RG process and on the instruments. To ensure consistent instrument information, specifically question wording and response options as well as their respective English translations, the structured **instrument documentation** finally comes into play. QuestionLink matches the harmonized instruments with the respective entry in the instrument documentation table to enrich the recoding documents. Users thus can look up pertinent instrument informations directly in the document. At this point, the QuestionLink process is complete. Making the recoding script documents available on the website currently is a manual process.

QuestionLink manual preparation workflow

While the QuestionLink engine undoubtedly saves a massive amount of work, it is still important to keep in mind that it only works once a well-structured response dataset has been compiled. In its current form, this means that we must identify all instruments which measure the same construct; and that across nine large survey programs and all their waves dating back to the eighties.

This involves several steps:

1. Conducting a prescreening whether a construct is measured in most of the QuestionLink surveys. This helps identify candidates for new harmonized constructs, because QuestionLink works best if constructs are measured often and in many surveys.
2. Based on the prescreening, the construct has to be clearly defined. This definition is key in selecting instruments. After all, OSE-RG can only harmonize instruments which measure the same construct.
3. Now codebooks, questionnaires and other survey documentation have to be searched to identify all potential instruments for the construct.
4. Depending on construct and the variability of the used instruments it might be necessary to empirically test if different instruments do measure the same construct. For the first few constructs, we sidestep this issue by choosing constructs which are measured with very similar wording across surveys (e.g., political interest or left-right orientation). However, in many cases deciding which instruments measure the same construct cannot be done at face value. Research on how to formalize this process is ongoing. However, for a quick

introduction, see this GESIS blogpost: Apples and Oranges: How to find out if two questions measure the same concept?

5. After the acceptable instruments have been identified, we screen for instruments which are the same across different surveys. In such cases, we treat it as a single instrument, albeit with data from different surveys.
6. Then we need to extract the relevant response data from the different surveys' scientific use files. During this process, we also need to extract survey design weights and other information. The result is a script which combines all those datapoints into a single, structured response dataset.
7. Lastly, we extract question wording and response option wording from the documentation. We then create the documentation table which will by the QuestionLink engine during the recoding script document creation stage.

Technical and methodological miscellanea

Finally, there are some specifics that we want to mention for transparencies sake, but which would have overcomplicated the explanations above.

Missing responses

There can be missing values in the response data that we gather from the different surveys. Missing values are potentially problematic, because OSE-RG may be biased by missing values. Specifically, OSE-RG will be biased if the respondents who refuse to answer an instrument measuring a construct differ in their true construct distribution. Then we can no longer assume that the response population for both instruments has the same true construct distribution (Kolen & Brennan, 2014). To make this less abstract, imagine the following case. In instrument A, very few respondents choose not to answer. In instrument B, however, especially respondents with a low construct expression (e.g., low life satisfaction) choose not to answer. This means that despite the two random samples, the true construct distribution in the remaining respondents now differs due to systematic dropout. In essence, this means if we apply OSE-RG despite a substantial number of missing values due to response omissions or refusal, we have to assume (or make a case) that the omission likelihood is unrelated to the construct expression.

Currently, we sidestep the issue in QuestionLink because (1) the surveys we include have low levels of response refusal in general, and (2) the constructs we chose are not particularly prone to response refusal. Consequently, missing values are rare and QuestionLink simply removes them before equating. If we want to include instruments with substantive response refusals or omissions in the future, however, then we need to examine this issue thoroughly.

Inverted response scales

Some instruments differ in the direction of their response scales. By direction we mean the order of response options. Some instruments start with responses which reflect low construct intensities and then options that represent higher construct intensities. Other instruments do the opposite.

QuestionLink solves this automatically by inverting response scores as necessary when harmonizing instruments with different directions.

Please note, however, that this is not the same as an inversion due to question wording. The numerical format can also be inverted if the question wording is negative or uses opposite expressions. In such cases, we plan to carefully examine if the negatively worded instruments still represent the same construct. The lexical opposite of a word does not have to be the psychological opposite: Mistrust may not be inverted Trust, but a different emotion altogether. With currently planned constructs, this issue does not yet present itself, but we are conducting methodological research on the issue for future constructs.

German reunification

Lastly, QuestionLink makes use of survey data dating further back than German reunification. This is relevant, because the survey we use exclusively sampled West-Germany before the reunification. In most cases, this does not interfere with the QuestionLink process. Whether we OSE-RG-match West-German samples with other West-German samples before the reunification or if we match Germany as a whole after the reunification should not matter as long as we assume that the instruments are similarly understood in both populations. The only issue that could arise is with time-relaxed links: Here the QuestionLink engine might harmonize a year before the reunification with one after the reunification, if we harmonize data from adjacent years. This would be a violation of the random groups design by matching a West-German sample with a unified German sample. However, this issue can only occur in some isolated instances and would most likely not make a difference in the final, aggregated recoding table. However, to be sure we added respondent level information on whether they live in the old or new federal states (i.e., former West or East Germany). This way, QuesionLink can now identify automatically if a wave is West German only or represents Germany as a whole. Links that would match across those different populations are identified and removed from the pool of valid links before applying OSE-RG.

References

Literature

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative Data Analysis in Clinical Psychology Research. *Annual Review of Clinical Psychology*, *9*(1), 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Menold, N., & Bogner, K. (2016). Design of Rating Scales in Questionnaires. *GESIS Survey Guidelines*. https://doi.org/10.15465/gesis-sg_en_015

Price, L. R. (2017). *Psychometric methods: Theory into practice*. The Guilford Press.

Schulz, S., & Weiß, B. (2014). *Harmonisierung und Synthese von paarbiografischen Daten: Ein Modellprojekt zur Verknüpfung von Forschungsdaten aus verschiedenen Infrastrukturen*.

Singh, R. K. (2020). Harmonizing Instruments with Equating. *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*, *6(1)*.

Slomczynski, K. M., & Tomescu-Dubrow, I. (2018). Basic Principles of Survey Data Recycling. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods* (pp. 937–962). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118884997.ch43

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.

Software

The QuestionLink Engine was written in R (R Core Team, 2021), using the RStudio IDE (RStudio Team, 2022). The QuestionLink Engine relies on packages from the tidyverse (Wickham et al., 2019). For parallel processing, we use the furrr package (Vaughan & Dancho, 2021).

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Vaughan D. & Dancho, M. (2021). furrr: Apply Mapping Functions in Parallel using Futures. R package version 0.2.3. https://CRAN.R-project.org/package=furrr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686